Bioschemas for Samples v0.1 [deprecated]

Summary

To deliver on the identified use cases for samples, we have identified a minimal set of properties to encapsulate identification, linking, and metadata descriptions. Some of these properties are existing standard schema.org properties, others require Bioschemas extensions. Table 1 outlines the minimal set of properties for the 'Sample' concept and Table 2 shows our recommendations for use of the 'PropertyValue' concept to markup additional characteristics of a sample described within a sample page. We also propose a new concept, 'Biomedical Code', which is a generalisation of the existing 'Medical Code' concept defined in the health-lifesci.schema.org extension.

Because we expect providers to be able to share partial information (to allow the complete picture to be assembled) all properties are optional in scope, there are no mandatory fields.

Sample

http://bioschemas.org/Sample [Proposed]

Property	Schema.org	Description	Type	Cardinality
identifier	http://schema.org/identifier	We expect this to be a BioSamples accession number	Text	0..1
name	http://schema.org/name	This is the identifier assigned to the sample by the Biobank.	Text	0..1/n?
description	http://schema.org/description	A description of the sample in free text. Ideally should not contain information that could be better expressed as key/value pairs	Text	0..1/n?
url	http://schema.org/url	An access URL for this sample, either in BioSamples or in a Biobank or elsewhere	URL	0..*
datasetUrl	https://bioschemas.org/datasetUrl [Proposed]	An access URL that provides a link to a dataset that contains data about this sample	URL	0..*
additionalProperty	http://schema.org/additionalProperty	A property-value pair representing an additional characteristics of the entity, e.g. "Organism: Homo sapiens" or "tissue type: leaf"	PropertyValue	0..*

Table 1: The 'Sample' Bioschemas concept and suggested properties

Property Value

http://schema.org/PropertyValue

Property	Schema.org	Description	Type
propertyId	http://schema.org/propertyID	The ID/name of the property (see schema.org definition). Ideally this is commonly used or qualified by namespace, but may be a local property name	Text
value	http://schema.org/value	The value of the property value node. It can be 'Text;', 'Number', 'Boolean', or 'StructuredValue'.	Text
code	http://schema.org/code	A biomedical code that describes the concept being captured by this key/value pair	BiomedicalCode

Table 2: The Property Value concept for schema.org and our recommended property uses

Biomedical Code

http://bioschema.org/BiomedicalCode [Proposed]

A generalisation of the specific concept MedicalCode (http://schema.org/MedicalCode, see https://health-lifesci.schema.org/MedicalCode). This concept can be adapted almost exactly as-is.

Note that this proposes a structured framework for expressing attributes of samples and how to describe ontology annotations of these attributes. It is desirable to ensure that this aligns with the P5: Phenotypes efforts.

Motivation

There are several competing efforts aimed at tackling the difficult problem of sample database interoperability. Doing so requires deep modelling and comprehensive validation against domain-specific standards, which are out-of-scope for a lightweight findability solution like Bioschemas. We therefore propose to avoid the issue of sample metadata modelling, and will not propose specific standards for e.g. expressing taxonomy of samples, instead providing a general framework for expressing metadata attributes ('key/value pairs'), that is compatible with current best-practice, including standards like MIAME, MIAPPE or MIABIS.

In conjunction with a general metadata framework, we will implement the Bioschemas framework to encourage data flow, data linking and reuse based on shared or synonymous identifiers. One key challenge for biobanks is to ensure that any experimental data generated on biological samples they provide is correctly linked back to the original record describing that sample. For example, experimental datasets should reference the sample name or identifier issued by the Biobank. If a scientist is attempting to reproduce results of an experiment, for example, they are likely to want to start with the same materials, and this may mean acquiring a tissue sample from the same source, where available.

The problem of identifying samples across multiple resources is complementary to the problem faced by public archives who wish to collect or report information on the provenance of samples. This problem is compounded by the issue of identifiability - often, metadata on samples ("sample sheets") are generated and submitted to public archives along with experimental metadata, and at this point samples used in experiments may be given a public identifier - for example, an accession in the BioSamples database. Biobanks have their own identifier system that may contain considerably richer metadata, some of which may be personally identifiable, confidential, or unconsented. It is unusual, in the absence of a dedicated project with a remit to support data co-ordination, for these identifiers to be reconciled.

Furthermore, whilst it is common for public data archives to provide API access, it is much less likely for Biobanks to do so. Biobanks are very likely, however, to have a website for ordering sample stocks from, and these pages will contain information on identifiers along with metadata attributes, probably expressed as key/value pairs. Providing an ability for public archives to harvest and potentially ingest this information, including linking to public data records, is highly desirable.

Proposed Solution

We propose to tackle the problem of sample identification and linking by providing a Bioschema mechanism for describing samples in terms of their known identifiers (specifically, public accessions vs biobank names), their links to datasets in the public domain, and a general framework for expressing metadata attributes.

Usecases

1. Biobanks should be able to crawl the BioSamples database to identify all the published (and searchable) datasets derived from samples they have provided
2. Public archives should be able to crawl Biobank websites, in order to identify samples that are known to have public accessions in the BioSamples database AND that can be made publicly available, and thereby link public samples to a provider ("where can I get more of this sample?").
3. In case of privacy or consent considerations, only the biobank should know what are the specific samples connected to publicly available datasets
4. Public archives should be able to crawl Biobank websites, in order to identify 'sanitised' sample metadata descriptions (again, in case of confidentiality or consent considerations). Biobanks remain responsible for ensuring only authorised metadata is visible, and can control access to restricted samples.

Assumptions

1. Each sample provided by a biobank has an opaque pseudo-anonymous identifier that is assigned by the biobank to identify a specific sample (referred to hereafter as the "sample name")
2. Each sample reported in a public archive or used to generate a public dataset has a public, BioSamples database accession (hereafter called "sample identifier").
3. In some cases, a biobank may issue different sample identifiers when providing the same sample to different projects. This may result in duplicated sample accessions in the BioSamples database

Given these use cases and assumptions, we will use Bioschemas to describe sample links. The main challenge is therefore the identification of links between sample identifiers (within Biobanks) and sample accessions (from the BioSamples database). This is not always possible without considerable additional curation effort, but of the 5 million samples in the BioSamples database, over 4 million declare either a 'synonym', 'sample source name' or 'source name' attribute, frequently used to encode the original biobank sample name. Exposing these in a structured manner through the BioSamples database would allow Biobanks to crawl and analyse this content, marrying sample that are recognised with their own internal identifiers.

Once this mapping is done, Biobanks can then re-expose these links through structured content on their own websites, allowing public resources to reciprocate links from public records back to the sample provider.

Implementation Study Outline

Objectives

- Facilitate the ingestion of sample metadata from data repositories (eg. Biobank databases) into registries like the BioSamples, BBMRI Biobank directory or the UKCRC Tissue Directory via Bioschemas.
- Engage and help data providers and developers of BioBank LIMS to test and adopt the exposure of sample metadata via Bioschemas
- Contribute to contextualise information from data sample registries (eg. BioSamples) and biobank sample repositories (eg. NL Biobank) and Biobank Registries (eg. BBMRI Biobank directory)
- Make registries like BioSamples compliant with Bioschemas.

Milestones

4.M1 Analysis and mapping of metadata already used in existing sample registries and defined by existing standards like MIABIS

4.M2 Define minimum information guideline based on the results of the mapping and feedback from registries of biological samples.

- Identify a minimum set of properties common across repositories

4.M3 Test adoption and improve specification with selected data repositories

4.M4 Propose any new suggested types or properties to schema.org

Deliverables

4.D1 Bioschemas specification

4.D2 Data repository using Bioschemas compliant markup

4.D3 Data registry using Bioschemas compliant markup

Example Implementation

Example snippet produced from https://www.ebi.ac.uk/biosamples/samples/SAMEA2340790

```
<div vocab="http://schema.org/" typeof="Sample">
  <h4 property="identifier">SAMEA2340790</h4>
    <table>
      <tbody>
        <tr>
          <td>Name</td>
          <td>ERS398461</td>
        </tr>
        <tr>
          <td>Description</td>
          <td property="description">84 Mixed species samples from ENA SRA</td>
        </tr>
        <tr>
          <td>synonym</td>
          <td property="name">S.lycLA4451_1</td>
        </tr>
        <tr property="additionalProperty" typeof="PropertyValue">
          <td property="propertyId">Organism</td>
          <td property="value">Solanum lycopersicum</td>
          <td property="code" typeof="BiomedicalCode">
            <a property="codeValue"
               href="http://purl.obolibrary.org/obo/NCBITaxon_4081">
               NCBITaxon_4081
             </a>
           </td>
        </tr>
        <tr>
          <td>External references</td>
          <td>
            <a property="datasetUrl"
               href="http://www.ebi.ac.uk/ena/data/view/ERS398461">
              <span property="name">ERS398461</span>
            </a>
          </td>
        </tr>
      </tbody>
    </table>
</div>
```

Additional Details

More associated information is presented in this working document: https://docs.google.com/spreadsheets/d/1NltFMzjqezpKhobmqEBEvASLziZjq73a_AqWnNi3IOs/edit#gid=0.